Research Discussion Paper No. 50

A Guide to the Unknown
Missing Data in Sample Surveys

by

David Odynak

Department of Sociology
The University of Alberta
Edmonton, Alberta

November 1987

ABSTRACT

The problems of missing data in sample surveys are examined in view of potential sources of missing data, detection, and the strategies available to handle missing data at various stages of the survey process. The discussion is limited to missing data on selected items in a survey excluding the issue of data missing as a result of nonresponse to the entire survey. Interestingly, the problems caused by missing data and specialized procedures to deal with missing data that are used in major surveys are generally not well known to students engaged in research in the social sciences. It is hoped that this paper will provide some information to students who construct and use survey materials in their research.

Introduction

Data missing in surveys are found in two forms: nonresponse to the entire survey (total nonresponse and/or noncoverage)and as nonresponse to selected items in the survey (item nonresponse) Also, responses in surveys made be found erroneous (content error) and in many cases treated as missing. In this paper, a number of aspects of both item nonresponse and content error in non-panel surveys are examined. The extent of nonresponse in a survey is a problem that can severely affect survey quality. Procedures used to handle missing response are one indicator of the level of technical sophistication employed in a given survey. Three main approaches for handling missing response can be adopted. First, careful planning and survey execution can help eliminate possible trouble areas. Second, item nonresponse and content error is dealt with during the data processing stage using editing and imputation. The most common approach, however, is to flag the missing data and pass the problem to the secondary data user. Several of the more widely used procedures across these three main approaches to handle missing data (item nonresponse/content error type) are highlighted in this paper.

Potential Sources of Missing Data

Item nonresponse can originate from several sources and are attributable to either the respondent or found in some stage of survey execution. Missing data on items are classified into two main types: intentionally excluded and unintentionally omitted responses. On the one hand, respondents are found missing on survey items because they were excluded by cost, redundancy, deemed inapplicable or not appropriate for answering the question. Intentionally omitted data from respondents frequently give rise to skip patterns in the data. Asking an unemployed

respondent what wage they make an hour for example would be a waste of time. Therefore, the unemployed are intentionally skipped over and the wage question asked of respondents who are working.

On the other hand, missing responses are unintentionally the result of insensitive or ambiguous questions. A respondent simply doesn't know a response to a question or, refuses to answer a question for a variety of reasons. Asking a respondent the amount of their personal income is one instance of a sensitive question that elicits a high percentage of nonresponse. Still another source of missing data may be attributed to a response deemed unacceptable because it is outside of the defined boundaries of acceptance--an outlier. For example, an 18 year female respondent legally married for 27 years is not a plausible set of values in many circumstances.

The potential source of missing data does not always lie with the respondent but could be the result of sloppy survey execution. Some possibilities for unreasonable values can be found in various stages of construction of the data set. Responses were being recorded improperly by an interviewer for example. Omissions or "blanks" may be found where values from the respondent should be. Coding can present problems where complex categories are involved. In addition, keypunching errors should be expected to occur during data transcription. Records can be lost or damaged during storage or in the data transfer stage. Most of these potential sources of error are generally not reported to secondary data users.

How Much is Too Much

Generally speaking a rule of thumb followed by some researchers is that under normal circumstances items with rates up to 10% missing can be used without serious concern. In the analysis of data from very small samples or from weak sample designs, a small percentage of missing cases can

easily distort relationships in terms of magnitude and direction. To make matters worse, some research shows that when cases are not missing at random, substantial biases in estimates can occur even with extremely high response rates (Anderson et al. 1983:480).

Problems caused by Missing Data

Two major questions surrounding missing data in a survey are first, what types of errors do missing data contribute to? and, when are missing data a problem?

Sampling-Nonsampling Errors Under Randomization Inference

Two types of errors are involved in survey sampling, sampling errors and nonsampling errors. Sampling errors involve the notions of repeated sampling and chance. A sample estimate could differ from the population parameter simply by the drawing of the sample. Another draw could produce a different sample estimate. If a number of samples are drawn under the same randomized conditions, the sample estimate will have as its expected value the population parameter. A sample estimate is unbiased when its average value is exactly equal to the population parameter (Cochran 1977:22). That is after repeated drawing of several samples the sample estimate does not deviate in a systematic fashion from the parameter.

Missing values (item nonresponse/content error) fall within the class of nonsampling errors. There are two types of non-sampling errors that occur, random error and bias (nonrandom error). When survey respondents are excluded on a nonrandom basis, the potential for selection bias exists (Berk 1983:390). When systematic bias is found evident the subset of missing respondents can no longer be considered a random subset of all respondents. When the subset of missing responses are nonrandom,making inferences to the population is generally no longer

appropriate.

Unfortunately, the majority of techniques for handling missing data rely exclusively on the assumption that the data are missing randomly. If the data are assumed missing randomly, one can go about infering that sample estimates produced for the respondents and nonrespondents differ only from the population by chance. It can be expected that random errors cancel out over repeated measurements (Fox and Tracy 1986:8).

In the case of nonrandom patterns of missing responses very little can be done. Extremely complex models based on assumptions made concerning the process causing the the missing response are rarely built. Usually, one is faced with the question of whether the bias is small enough to proceed. A sizeable proportion of nonresponse makes it virtually impossible to assign confidence limits to estimates from survey results. Essentially, with a sizeable proportion of nonresponse one is left in the position of guessing what the size of the bias is without data to substantiate the guess (Cochran 1977: 361). Nonrandom patterns of missing data frequently occur when survey respondents are intentionally excluded from the sample. Analysis of the remaining respondents after routing or skipping in surveys is a tricky excercise. Sometimes data from skipping are dealt with via procedures for censored and truncated distributions of variables (See Berk and Ray 1982;Berk 1983).

Missing Respondents Not the Same as Respondents

Cochran (1977) recommends thinking of nonrespondents as a separate stratum. In total nonresponse to a survey, it has been found repeatedly that the nonresponding stratum differs on several characteristics from the units that were measureable. When considering item nonresponse we might consider treating the different reasons for nonresponse as separate strata.

Clearly, a missing value resulting from a office or clerical error is different than refusal to answer a survey question. However, there has been a tendency to collapse all the reasons for nonresponse as three, two or even one category. Many of the techniques to adjust for nonresponse fail to distinguish between the different reasons for missing. Instead the different reasons for nonresponse are treated as one category of missing and the same rules generating values to replace the missing value are applied to all missing cases.

Loss of Data

Another problem caused by missing data is the potential for loss of data, especially in multivariate analyses. If one is dropping incomplete records due to missing data from the analysis the problem of too few cases on several variables may arise. Interestingly, for a small survey, a high number of missing responses on several questions may render any multivariate analysis virtually meaningless. For example, if 5% of the responses are missing randomly on each of 5 variables in a multivariate analyses a potential 25% of the sample could be dropped. The adjustments for the calculation of estimates from samples with missing data can be very complex. Many secondary data users are unaware of the necessary adjustments required for incomplete sample observations.

Embarrassment

Related to the loss of data issue is credibility. Census organizations and major survey research organizations are subject to public release of their data sets and results. The number of missing cases generated by item nonresponse and total nonresponse is one objective measure of the quality of survey data. While many realize that inconsistencies and unknowns contribute to missing values, many users are not as sophisticated. When people discover anomalies and "not stated" columns in tabular data

some degree of credibility on behalf of the agency is questioned (Banister 1980:17).

Worry over the extent of item missing response as a problem depends primarily on the type of survey and organization conducting the survey. For a census type survey, rarely are missing data tolerated. Some survey organizations will give all cases a response for every item in a survey. Great planning and effort is expended ahead of time so that anticipated missing data is dealt with during survey execution and data processing. However for many researchers facing the problem of missing data, building elaborate models ahead of time may not be as an important a priority as turning out the results on time.

Detection of Missing Data Problems

Frequency Distribution

Initially, to assess the extent of missing response for an item, a frequency distribution of responses can be used. If special codes exist for missing values (i.e. don't know, not applicable, or refusal) most statistical packages can generate a table showing the proportion of these nonresponses to valid responses. In addition, respondents with values outside of the accepted range,the outliers,will also show up in a frequency distribution if not already designated as missing. Usually, special flags for missing values are used in the statistical computing packages so that missing values do not contribute to the estimates.

Crosstabulation

Crosstabulation of respondents with missing values for one variable x with values for another variable y reveals the distribution of missing cases for values for the second variable (y). If the values for the missing respondents on x appear to be different on y than the values for nonmissing

respondents on x, then the pattern of non response is likely nonrandom. In addition, where values are missing on both x and y, crosstabulation can show the degree of overlap between variables on nonresponse. In other words, a respondent who refuses to answer item x, may have also declined to answer item y.

Dummy Variable Analyses

Cohen and Cohen (1975) suggest the use of dummy variable analysis when the data are suspected to represent a non-random subsample of the survey population. In this procedure a dependent variable Y is regressed on the dummy variable X representing the variable with missing values. Nominal level data can be used in the regression if the data are collapsed into two values. The collapsed nominal variable is treated as an interval level variable with just two categories in the regression. (Lewis-Beck 1980:67).

Dummy variable analysis assigns constants to both missing and non-missing data. Typically, 1's and 0's are used, where X=0 would indicate the presence of an attribute in the variable X and X=1 an absence. The Y intercept represents the mean value of respondents (as a reference group) and the slope is the difference between the conditional mean Y values of the missing respondents (when X=1) and nonmissing respondents (when X=0). If the cases with missing values have a different Y mean than the respondents, then the statistical significance of the difference can be tested using a t-test on the slope. When the difference is shown to be statistically significant, then the missing data on X are not random.

The analysis of missing data using dummy variable regression can be extended by using other constant (plugged) values for missing respondents. For example, the mean of available scores can be plugged in where the data are missing.[1] Cohen and Cohen (1975) also suggest the use of alternate

coding strategies and non-linear transformations to examine for the non-random presence of missing values. Many of these extensions to the basic regression equation can incorporate a researcher's theoretical expectations on the behaviour of missing respondents vis-a-vis respondents.

The techniques suggested by Cohen and Cohen have several advantages. One major advantage is that records with missing data are not thrown out of the analysis. Information on missing cases is used in a regression analysis with a complete data set. In addition, the dummy variable regression technique is suitable for both nominal and categoric data.

Strategies For Handling Missing Data

Several strategies have been suggested for dealing with the problem of missing data. Careful survey execution and design serves to minimize the number of missing cases on items. A return to the field in order to complete missing information from the respondents is practiced when feasible financially and from the standpoint of timeliness. In addition to extensive field editing, Statistics Canada, and other major survey centres, prefer to impute the missing values as part of the data cleaning procedures. These agencies have access to information that is not usually available to the public for confidentiality reasons. Therefore, imputation in the data processing stage may develop more efficient imputations than the users (David et al. 1986:29)[2] Others prefer to leave the missing data alone, flag it and pass the problems on to the data analyst. All of these paths have relative advantages and disadvantages in various situations.

Careful Survey Execution

Careful planning in the design and execution serves to minimize the number of missing responses in a survey. A variety of helpful strategies

exist apart from exerting more care and attention in the coding, editing and keypunching phases. In some major census survey operations, respondents are required by law to answer the questions. To help cut losses due to nonresponse, a mandate or some air of officiality to convince the respondents to take the survey seriously may be needed. Some assurance of the degree of confidentiality and anonymity for the respondents answers through informed consent is a major ethical issue in survey research (See Dalenius 1983). Many survey organizations send out cover letters informing the potential respondents of their right to refusal and extend the offer of confidentiality to all personal data. Government agencies go to extraordinary lengths to protect the confidentiality of information. At Statistics Canada sensitive personal data are often rounded and data suppressed from certain cells so that the individual cannot be identified outside of the agency.

A pretest of the questionnaire flags potential problem questions. Some questions may have to be redesigned and require the addition of similar questions to supply more information on the missing respondents. The randomized response technique has also been used to encourage more response on sensitive questions.[3] Unless some information on variables with missing items are available from other surveys, a pretest may in fact be the only way to go in establishing a series of models for handling missing data during the collection phase.

Going Back to the Field

When feasible in terms of cost and time, another alternative is to go back to the respondent and recollect the information that were found to be missing. Field editing during and after data collection helps clear up anomalies, blanks, keypunch errors and much of the misunderstanding between interviewer and respondent that jeopardizes the overall quality of the data

coding strategies and non-linear transformations to examine for the non-random presence of missing values. Many of these extensions to the basic regression equation can incorporate a researcher's theoretical expectations on the behaviour of missing respondents vis-a-vis respondents.

The techniques suggested by Cohen and Cohen have several advantages. One major advantage is that records with missing data are not thrown out of the analysis. Information on missing cases is used in a regression analysis with a complete data set. In addition, the dummy variable regression technique is suitable for both nominal and categoric data.

Strategies For Handling Missing Data

Several strategies have been suggested for dealing with the problem of missing data. Careful survey execution and design serves to minimize the number of missing cases on items. A return to the field in order to complete missing information from the respondents is practiced when feasible financially and from the standpoint of timeliness. In addition to extensive field editing, Statistics Canada, and other major survey centres, prefer to impute the missing values as part of the data cleaning procedures. These agencies have access to information that is not usually available to the public for confidentiality reasons. Therefore, imputation in the data processing stage may develop more efficient imputations than the users (David et al. 1986:29)[2] Others prefer to leave the missing data alone, flag it and pass the problems on to the data analyst. All of these paths have relative advantages and disadvantages in various situations.

Careful Survey Execution

Careful planning in the design and execution serves to minimize the number of missing responses in a survey. A variety of helpful strategies

exist apart from exerting more care and attention in the coding, editing and keypunching phases. In some major census survey operations, respondents are required by law to answer the questions. To help cut losses due to nonresponse, a mandate or some air of officiality to convince the respondents to take the survey seriously may be needed. Some assurance of the degree of confidentiality and anonymity for the respondents answers through informed consent is a major ethical issue in survey research (See Dalenius 1983). Many survey organizations send out cover letters informing the potential respondents of their right to refusal and extend the offer of confidentiality to all personal data. Government agencies go to extraordinary lengths to protect the confidentiality of information. At Statistics Canada sensitive personal data are often rounded and data suppressed from certain cells so that the individual cannot be identified outside of the agency.

A pretest of the questionnaire flags potential problem questions. Some questions may have to be redesigned and require the addition of similar questions to supply more information on the missing respondents. The randomized response technique has also been used to encourage more response on sensitive questions.[3] Unless some information on variables with missing items are available from other surveys, a pretest may in fact be the only way to go in establishing a series of models for handling missing data during the collection phase.

Going Back to the Field

When feasible in terms of cost and time, another alternative is to go back to the respondent and recollect the information that were found to be missing. Field editing during and after data collection helps clear up anomalies, blanks, keypunch errors and much of the misunderstanding between interviewer and respondent that jeopardizes the overall quality of the data

and future surveys.

Despite careful survey planning and attempts at field editing, the need for a strategy to handle missing data at another stage of the processing or when actually conducting data analysis might arise. Keypunching errors are often rectified through verification, where the data set is punched in twice and compared. In many survey organizations the resolution of errors is frequently handled in the batch mode by the data collectors (Karwei and Meyers, 1983:396)[4] The strategies for handling missing data by data collectors discussed here are: using redundant questions, editing and imputation during the data cleaning stage.

Edit and Impute

Editing and imputing procedures are commonly found in census and major survey organizations. Essentially, editing refers to the process that checks the validity of responses in the questionnaire and any subsequent changes made to the responses by data collectors. The edit procedure attempts to remove discrepancies in the data that are systematic rather than random (Pullum et al. 1986: 312). For example, if one particular interviewer continually misses a portion of the survey when recording the responses then manual editing through a quick review of all questionairres by trained personnel could serve to detect and help correct this systematic bias. Typically, either the respondent and interviewer would be contacted again or some of the data discarded.

Imputation refers to the filling in of unacceptable responses with more plausible values. A major argument for imputation is that filling the values is better than doing nothing at all. Doing nothing at all assumes that nonrespondents are only randomly different than respondents. By narrowing the scope of nonresponse using selected characteristics the imputation is not

chosen at random but makes use of selection within carefully designated cells. Quite often, editing and imputing are done close together with the imputed answers required to pass the edits before entry into the database. Felligi and Holt (1976) present a generalized system of edit and imputation followed by several statistical agencies. In their generalized system, the edit-imputation procedure consists of identifying the smallest set of variables whose values could be changed so that the resulting record will satisfy all edits and then imputing values for those fields that will maintain the distribution of data as represented by the records that passed the edits (Karweit and Meyers 1983:396). Little and Smith (1987) combine various statistical methodologies in estimation, outlier detection and multivariate analysis and apply these methods to the edit and imputation of survey data.

Hot and Cold-Deck Imputation Procedures

Imputation of responses for missing data can be carried out in various ways. Two main techniques used by major survey organizations and census bureaus are the hot- and cold-deck procedures to handle missing data. These deck procedures are named after the punched card reading techniques used by researchers when computers used IBM punch cards. Both deck procedures require some prior knowledge of nonresponse from previous surveys. Cold deck procedures refer to situations when the answer used to fill in the missing blanks come from another source outside of the survey. Hence, in computer card reading facilities a new deck of cards would have to be read in. Some of the procedures in use include the use of historic data from previously conducted surveys, rule of thumb values deemed reasonable by the analyst (ad hoc values), and matched or proxy variables from other surveys where linking is possible (Sande 1982). One example of the cold-deck procedures are the best guess solutions for imputing values for dates in the World Fertility

Surveys. In these cases, where no other information was available to help correct missing information on dates, an informed guess would be made (Pullum et al. 1986).

Hot-deck refers to imputation procedures where the information is taken from the same card deck that is being read in. Values missing on various items are replaced by values from "donor" records. The donors can be selected in a number of ways. Frequently, the variables to be imputed are selected before data processing and "imputation classes" set up so that the number of potential donors are limited by a few distinguishing characteristics. The donors are then chosen either randomly, randomly within the imputation classes or as a result of being the last record with the same characteristics. Several variations of the hot-deck procedure have emerged with the new found capabilities of high speed computer data processing. Sande (1982) and Kalton and Kasprzyk (1986) provide descriptive accounts of some of the hot-deck methods used in major surveys.

One major strength of the hot-deck procedures is that feasible values are provided for both nominal and categoric variables. In other words, values such as 2.5 children are not used for imputation. These techniques are suited toward the general manipulation of several variables in the data set and make use of the knowledge of the data set. Major survey organizations try to utilize the joint distribution of variables by imputing several missing items using one donor in an attempt to preserve the covariance structure of the data. All procedures require some modelling process of how nonrespondents act vis-a-vis respondents.

Some disadvantages of the hot-deck procedures are that the variables used to form imputation classes must be strongly correlated with the survey variable or the procedure is not much of an improvement (Ford 1983). The

hot-deck procedures mentioned here assume that after the sorting variables are taken into account the non-responding population are distributed randomly (Kalton 1986:14). Although the population totals and averages can be estimated using standard complete data formulas, the estimation of variances is a different story (Dempster and Rubin 1983:8).[5] Variances of estimates using the hot-deck model are typically underestimates and have to be adjusted accordingly.

Redundant Questions

Where one question is found to have a sizeable portion of missing cases, one solution has been to use other items to construct a response for the missing. For example, family income is theorized to consist of husband's income+wife's income+income from other sources. While providing a response for the missing item, such a technique may also show that non-respondents are different in their income attainment than complete items. One can also construct family income from other sources and compare the result with actual reported family income for respondents answering all questions. This technique of imputing values of respondents using reconstruction depends on careful planning in the survey design stage. Moreover, prior knowledge of trouble variables either from the literature or a pretest is typically required for such an exercise.

Evaluation

One of the more difficult decisions facing either the data collector or the secondary analyst is whether or not the strategy used for missing data makes a difference. Sometimes, blind deletion or improper imputation can actually worsen the situation. Usually the evaluation of different techniques rests on the assumption that the data are missing randomly. Therefore, the tests normally consist of artifically eliminating cases on

a random basis from the data set and comparing the estimates produced by the various missing data procedures with the estimates using the complete data set. Unfortunately, there is a lack of standardization between the missing data conditions for the evaluation of different techniques to handle missing data. Evaluations of procedures are extremely dependent on the sample size used, the variables under study and the magnitude of missing cases. Moreover, there is no guarantee that the missing data routine will produce a similar good fit for nonrespondents under less artificial conditions.

Sometimes, imputations for nonrespondents on one survey can be assessed through information from other surveys. Similar to the evaluation of population estimates, the various procedures for imputing data can be compared to a standard. Estimates of population are compared to later census figures obtained for the same population. In the case of item nonresponse models the respondent's actual value from another source can be linked to the item missing on a survey. For example, David et al. (1986) were able to compare imputations for income in the Current Population Survey in the United States with the respondent's income reported in Internal Revenue Service data. An exact match was made using social security numbers. However, access to such confidential data is rarely available to examine imputation models for random and nonrandom patterns of missing data.

Multiple Imputation

Most of the imputation strategies ignore the fact that the missing values are not known (Herzog and Rubin 1983). Single imputations for missing data generally produce underestimated standard errors. One interesting approach to the bundle of problems associated with filling in the missing data with only one value is to provide multiple imputations

(Herzog and Rubin 1983;Rubin 1986). Under the proposal of multiple imputation, two or more imputations are generated for each missing value, estimates are produced using a new data set for each imputed value. Then the estimates are averaged and the result is one valid inference for each model. The idea of multiple imputation is attractive in that it can provide an appropriate mechanism for making inferences to the population. In addition, it allows the researcher to test the sensitivity of the estimates to the values imputed from various models, so called "imputation error" (Andersen et al. 1983:480; Kalton and Kasprzyk 1986:13). The major problem inhibiting the production of multiple imputations as a common practise are the enormous computing costs and time involved to produce the multiple data sets required.

Major Concerns Over Imputation

While imputation at the data processing stage has certain advantages over the adjustment for missing data at the time of data analysis, some criticism and many recommendations have surfaced over the various procedures to handle missing data. Many secondary data users are unaware that imputation has taken place. When non-response has been rectified through imputation at the data stage, users generally analyse the filled-in data as if it were complete; ignoring the imputation (David et al. 1986:29). Recommendations that two sets of data be provided one with flagged missing and the other with imputation usually means a lot of work and money. Sometimes imputation procedures are generally outlined in tables and special publications with numbers of nonresponse provided and definition of imputation methods used (See Boxhill 1986 as an example). Unfortunately, many major data sets contain imputations but the discussion of imputation is limited to a footnote. Users interested in more information may have to

contact several sources to ascertain the degree and method of imputation performed.

Another major concern with the handling of missing data, is the lack of attention paid to the different types of reasons why the data are missing. There is an overriding tendency to handle the various reasons for item non-response as just one category missing with no distinction between those who refuse, don't know, eliminated on purpose or accidently from the original sample. Hopefully, for many researchers when all the various reasons are mixed together for item nonresponse the result is a nonrandom pattern. However to conduct an indepth analysis of different causes of missing data producing different patterns of nonresponse represents a great deal of work in relatively uncharted territory.

Alternately, a less favorable view on imputation has emerged. One can view edit and imputation as one of many alterations to the data that users are generally unaware of. Thus, edit and imputation procedures fall into the same class as rounding for confidentiality reasons, the handling of multiple response and other data alteration techniques that make sense to the survey organization but may impair outside analyses. Blind or semi-informed imputation adopted to make the data look attractive, and the organization look modern and efficient can lead to extremely distorted results (Banister 1980). At the extreme end is the view that if the practise of imputation is accepted then there is no real need to do the survey in the first place (Bateson 1984).

Past the Data Collection Stage

Usually in the social sciences, data are collected and processed and much of the trouble of missing variables are left to the secondary data analyst. The hot-deck and cold-deck imputation procedures during the edit

stage are unavailable or not a primary concern. Quite often, the secondary data user is faced with missing items in a survey that he/she had no active role in constructing. Many of the options available to the social scientist to deal with missing data depend on theoretical expectations, methodological expertise and what the current statistical computing package offer them. Some of the more popular options are reviewed here.

Flagging Missing Data

Certain rules should be followed when declaring missing values. Missing values flagged by certain codes, usually are kept separate from the range of accepted values. Chuck Humphrey, data analyst at the University of Alberta, suggests that codes be chosen which are contiguous to the legitimate response set to minimize the amount of memory reserved for the data array. For example, if the legitimate range of values are 1 thru 5, using 9's and 99's, a common practise,can lead to waste of computer storage and funds especially when done over a large range of variables. Instead he proposes that contiguous numerical codes such as 0's or 6's be used to flag missing data. Another useful suggestion for coding missing data is to try to use the same missing value for as many variables as possible. When using one missing value, such as 6's for all the variables in the above example, one can make use of global statements within the statistical computing packages to save time and reduce potential confusion.[6]

Most of the current options for handling item nonresponse are strictly dependent on one major assumption, that is the data are missing at random under various conditions. Usually the major definitions are that: first, the pattern of missing elements for a variable are distributed randomly within that variable. That is values are not missing because a value is too sensitive to report or a range of values were not coded properly. Second,

missing data in one variable are not related to the presence or absence of certain values on another variable. These values are systematically linked to a range of values on another variable or missing information on another variable. Extension and illustration of the conditions for randomly produced missing data are provided in Andersen et al. (1983) and Kim and Curry (1977).

Listwise Deletion and Item Deletion

The most selected option for handling missing data is to leave the data flagged and drop the entire record from the analysis. The calculation of subsequent estimates are based on a sub-sample of the original respondents. Deletion across several variables (listwise deletion) in a multivariate analysis can severely limit the ultimate number of respondents used in the production of estimates. The higher the number of records deleted, the less satisfying the assumption that the remaining data represent a random sub-sample of the full sample (Hertel 1976:461). It is the default mode of many of the computer package statistical routines. When researchers are faced with unrealistically small samples after listwise deletion of records across items , one alternative has been to drop entire items from the multivariate analysis. Deletion of entire items is an attractive option when the data are missing at random and the number of records deleted overall will be relatively small.

Pairwise Deletion

Pairwise deletion is an option available in several packages that considers only two variables at a time for deletion of items with missing values. For example, in the calculation of a correlation matrix using several variables, only values missing on a given pair of variables are deleted from the estimation of the coefficients. The benefits of using the procedure in multivariate analyses are that there is less waste of data and the expected values can be unbiased. However, the interpretation of

multivariate analyses such as multiple regression can be difficult since different sample sizes are reported for each coefficent. Also, the possibility of some impossible values being generated exists when the same cases are not used to estimate different coefficients (Norusis 1983:95). Berk (1983b:540) suggests that when rates of missing data are high, it may be w to pay the price of inefficiency with listwise deletion than try to interpret estimates from the pairwise deletion procedures. If time and cost permit a good strategy usually involves computing two sets of estimates, using both pairwise and listwise deletion, and comparing the results.

More Imputation

Besides deletion of records with nonresponse, several imputation procedures to estimate missing data are available for the secondary data user. Like the hot-deck cold procedures used in data processing, the methods are based on models built on assumptions concerning the distribution of missing values in the data set. The more popular methods of imputation reviewed here are mean substitution, use of the group mean for missing cases, and linear regression to predict the missing values.

Mean Substitution

Mean substitution appears as an attractive option to handle missing values available in many of the current computing statistical packages routines. The mean based on complete cases is calculated and this value is imputed to the missing cases. The major assumption being that the missing respondents are really the typical case. However, using the mean to represent missing cases in computing estimates can lead to biased estimates. A bad quality associated with using the mean is that the variance is reduced and subsequently other measures such as t correlation coefficient are conservative (Hertel 1976).

Class Means

In the class means solution various sort or auxiliary variables are selected and the sample is broken down in to various strata. Within the various strata a mean is assigned to cases where the values are missing. The major assumption, like the random assignment and sequential hot-deck procedures, being that the missing cases are randomly distributed within the newly formed strata. The sort variables should be strongly correlated with the variables with values missing (Anderson et al. 1983). Also, the sort variables should not be strongly related to each other. The major advantage of the group mean over the overall mean imputation scheme are first that th stratification takes into account additional information for estimation. Second, the group mean procedure takes into account the variation between groups or strata (Kalton 1983:6).

The major disadvantages of using group means to impute missing values are that spikes appear in the overall distribution of the variable after imputation. Also, there is an absence of within strata variation that will affect the estimates (Sande 1982:148).

Regression

Imputation of missing values using linear regression is one of the more powerful techniques available to the secondary data user. The method is suitable for estimating missing values for one or two variables but is generally not recommended for imputing all the variables in the data set. Regression techniques, are similar to the hot-deck procedures in using auxiliary information in the data at hand for imputation. Under the simplest situation missing values are predicted from a set of independent variables. The regression estimates replace the missing values. For example, missing personal income (dependent variable) is frequently predicted in many surveys from respondent characteristics such

as age, sex, education, and occupation. Missing and non-missing responses for income are assumed to follow similar distributions. After the variation in income is explained by the various predictors, the remaining unexplained variance is assumed to follow a random pattern.

Several assumptions apply when using the linear regression equation. The assumptions usually are that the expected value of the errors is zero, the errors are normally distributed and the errors are statistically independent of the predictor variables (Blalock 1979:387).

Several extensions to the regression procedure to impute missing values have been suggested in the literature. Complex transformations for non-linear regression can be done but the models have to be specified. Regression estimates can also have a residual added that is obtained from a hot-deck from the actual residuals of fitted values or randomly generated from an expected distribution (Sande 1982:150). The chief advantage being that the resulting imputation produces a more feasible value similar to a hot-deck imputation (Kalton and Kasprzyk 1986:10).

What's Available

Imputation and other strategies for handling missing data at the data analysis stage is sometimes dictated by the availability and features of major software programs. Most major statistical computing packages offer special flags for handling missing data so that the missing data do not enter the calculation of estimates. Multivariate procedures in major packages such as SPSSX provide both listwise and pairwise deletion of cases. Mean substitution for missing data in regression is available. Probably, the most detailed missing data procedure available to the secondary data analyst is the program PAM under the BMDP statistical package. PAM provides various descriptive tools to help analyze patterns of missing data. In

addition, the program PAM can estimate missing values using regression under several options: forced entry, maximum likelihood and stepwise routines.

Summary

Missing values in sample survey data can arise from several sources attributable to either the respondent or found as some feature of survey execution. The worst scenario associated with missing data is that the data are not random, bias is present and inferences to the population are no longer appropriate. Detection and resolution of errors can take place in various stages of the survey process. Careful planning and extensive field editing have the potential to eliminate major sources survey error. Major survey centres, like Statistics Canada, prefer to impute values for missing data during data cleansing or processing. These elaborate techniques are modelled and evaluated with resources normally outside the range of the secondary data user. There are some criticisms with imputing missing data at any stage of survey research. Some of the criticisms have been answered by providing the secondary data user with more information on the imputation procedure, degree of imputation and rules used to alter the data. Hopefully, some of these elaborate editing and imputation techniques will be made more accessible to smaller survey centres and secondary data users. Of course another alternative is to flag missing data and pass the problem over to the secondary data user.

Many secondary data users are unaware that their data sets have been subjected to imputation. Usually, imputation is done under rigid conditions so that joint distributions of variables are maintained and imputations pass stringent editing rules. But when missing data are flagged and passed on to the secondary data user there are several models for the user to explore. Generally, choices are usually limited to what options the major statistical

computing packages offer. Most are based on the assumption that the data are missing at random. It is expected that after more procedures and techniques for edit and imputation are made readily available, secondary users will have more options and information to deal with their potential problems of missing data.

ACKNOWLEDGEMENTS

This paper is a condensed version of a specialization paper written under the supervision of Professors Karol Krótki, P. Krishnan, Frank Trovato and George Jarvis. The author is indebted for their insightful suggestions, revisions and lively discussion concerning this controversial topic in the social sciences.

Notes

[1]Plugging in the mean of the responding cases as a constant to represent the missing respondents can simplify the analysis and cut down on the number of computing steps. Using the mean to represent missing cases in the dummy variable regression turns partial slopes into simple slopes. Note, however, that the mean substitution option in regression many of the statistical computing packages do not accomplish the above.

[2] Data collectors typically have access to information that is confidential and not available to secondary data users to sort, classify and build models for imputation. In addition, data collectors may also be privy to other confidential data sources for outright imputation or evaluation of imputation models to match imputed data with the actual respondent value.

[3] See Dalenius (1983) for a discussion of some of the issues concerning informed consent in survey research.

[4]Randomized Response Technique (RRT) is a survey technique used to ensure the anonymity of the respondents while providing some estimate of the prevalenc description of the basic RRT and recent enhancements suggested. Emrich (1983) provides some illustrations of the use of RRT for sensitive questions in areas such as abortion and child abuse.

[5] Under the batch mode, data processing is set up as a series of steps classified as a job. Each step is executed sequentially and if one step fails the entire job aborts and the other steps are not executed. Typically, the user has no control over the execution of the steps after submitting the job. Jobs may be lined in a queue awaiting execution. The alternative to the batch mode is to run the job interactively. The advantage being that the user usually recieves the results promptly and can easily modify and execute the various steps at their convenience.

[6]See Ford (1983) and Herzog and Rubin (1983) for some of the formulas for estimating the variances under the hot-deck procedures. Most of the formulae provided follow the assumption of simple random sampling.

[7]In SPSSX, if all the variables have the same value designated as missing one can use the statement missing all (value). However, lumping all variable missing values under one code can easily blurr the distinction between don't know, not applicable or refusal to answer.

[8]Frequently, the logarithm of the reported value of income is used as the dependent variable in the regression.

References

Anderson, Andy,B., Alexander Basilevsky and Derek P.J. Hum. 1983. "Missing data: A review of the literature." Pp. 415-494 in Peter Rossi, James Wright and Andy B. Anderson (eds.), Handbook of Survey Research. Orlando, Florida: Academic Press.

Banister,Judith. 1980. "Use and abuse of census editing and imputation." Asian and Pacific Census Forum, East-West Population Institute 6(3).

Bateson, Nicolas. 1984. Data Construction in Social Surveys. London: George Allen and Unwin.

Berk, Richard. 1983a. "An introduction to sample selection bias in sociological data." American Sociological Review 48 (4):386-398.

Berk, Richard. 1983b. "Applications of the general linear model to survey data." Pp. 495-546, in Peter Rossi, James Wright and Andy B. Anderson (eds.), Handbook of Survey Research. Orlando, Florida: Academic Press.

Berk, Richard and Subhash C. Ray. 1982. "Sample selection bias in sociological data." Social Science Research 11(4): 352-398.

Blalock, Hubert M. Jr. 1979. Social Statistics. Second Edition. Toronto:McGraw-Hill.

Boxhill, Walton, O. 1986. A User's Guide to 1981 Census Data on Place of Birth, Citizenship and Immigration. Statistics Canada Occasional Paper Catalogue 99-946. Ottawa: Ministry of Supply and Services.

Cochran, William G. 1977. Sampling Techniques. 3rd edition. Toronto: John Wiley and Sons.

Cohen, Jacob and Patricia Cohen. 1975. Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences. Hillsdale, N.J.:Lawrence Erlbaum.

Dalenius, Tore. 1983. "Informed consent or R.S.V.P." Pp. 85-106 in William G. Madow, Ingram Olkin and Donald Rubin (eds.), Incomplete Data in Sample Surveys. Volume 2 Part IV. Toronto: Academic Press. 85-106.

David, Martin, Roderick J. Little, Michael Samuhel and Robert K. Triest. 1986. "Alternative methods for CPS income imputation." Journal of the American Statistical Association 81(393): 29-41.

Dempster, Arthur and Donald Rubin. 1983. "Introduction." Pp.3-10 in William G. Madow, Ingram Olkin and Donald Rubin (eds.), Incomplete Data in Sample Surveys. Volume 2 Part IV. Toronto: Academic Press.

Dixon, W.J. (Ed.). 1981. BMDP Statistical Software 1981. Berkeley: University of California Press.

Emrich, Lawrence. 1983. "Randomized response techniques." Pp.73-80 in William G. Madow, Ingram Olkin and Donald Rubin (eds.), Incomplete Data in Sample Surveys. Volume 2 Part II. Toronto: Academic Press.

Felligi, Ivan. and D. Holt. 1976. "A systematic approach to automatic edit and imputation." Journal of the American Statistical Association 71(353): 17-35.

Ford, Barry. 1983. "An overview of hot-deck procedures." Pp.185-207 William G. Madow, Ingram Olkin and Donald Rubin (eds.), Incomplete Data in Sample Surveys. Volume 2 Part IV. Toronto: Academic Press.

Fox, James and Paul E. Tracy. 1986. Randomized Response Techniques. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-001. Beverly Hills: Sage Publications.

Hartley, H.O. 1980. "Statistics as a science and as a profession." Journal of the American Statistical Association 75(369): 1-7.

Hertel,Bradley. 1976."Minimizing error variance introduced by missing data routines in survey analysis." Sociological Methods and Research 4 (4): 459-474.

Herzog, Thomas and Donald Rubin. 1983. "Using multiple imputations to handle nonresponse in sample surveys." Pp.210-245 in William G. Madow, Ingram Olkin and Donald Rubin (eds.), Incomplete Data in Sample Surveys. Volume 2 Part IV. Toronto: Academic Press.

Kalton, Graham. 1983. Compensating for Missing Survey Data. Ann Arbor, Michigan: Survey Research Center, Institute of Social Research, The University of Michigan.

Kalton, Graham and Daniel Kasprzyk. 1986. "The treatment of missing data." Survey Methodology. 12(1): 1-16.

Karweit,Nancy and Edmund Meyers,Jr. 1983. "Computers in Survey Research." Pp.379-414 in Peter Rossi, James Wright and Andy B. Anderson (eds.), Handbook of Survey Research. Orlando,Florida: Academic Press.

Kim, Jae-on and J. Curry. 1977. "The treatment of missing data in multivariate analysis." Sociological Methods and Research 6(2): 215-240.

Lewis-Beck, Michael. 1980. Applied Regression: An Introduction. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills: Sage Publications.

Lillard, Lee, James Smith and Finis Welch. 1986. "What do we really know about wages? The importance of nonreporting and census imputation." Journal of Political Economy 94(3):489-506.

Little, Roderick and Philip Smith. 1987. "Editing and imputation of quantitative survey data." Journal of the American Statistical Association 82(397):58-68.

Norusis, Marija J.1983. Introductory Statistics Guide SPSSX. Toronto: McGraw-Hill.

Platek, R. and G.B. Gray. 1978. "Non-response and imputation." Survey Methodology 4(2):144-176.

Platek, R. and G.B. Gray. 1986. "On the definitions of response rates." Survey Methodology 12(1):17-27.

Pullum, Thomas, Trudy Harpham and Nuri Ozsever. 1986. "The machine editing of large-sample surveys: The experience of the World Fertility Survey." International Statistical Review 54(3): 311-326.

Rubin, Donald. 1986. "Basic ideas of multiple imputation for nonresponse." Survey Methodology 12(1):37-47.

Sande, Innis.G. 1982. "Imputation in surveys: Coping with reality." The American Statistician 36(3.1):145-152.

RESEARCH DISCUSSION PAPER SERIES

NO.

39. The Relationship Between Migration and the Provincial Divorce Rate in Canada, 1971 and 1978: A Reassessment.
Frank Trovato. February 1985.

40. Interprovincial Migration and Suicide in Canada, 1971 - 1978
Frank Trovato. May 1985.

41. Economic Status of Immigrant Women at Age 30, Canada, 1981.
Carl F. Grindstaff and Frank Trovato. June 1985.

42. The Effects of Marital Dissolution on Suicide: The Canadian Case.
Frank Trovato. June 1985.

43. Fertility Estimates for Bangladesh Based on Birth History Data.
S.M. Shafiqul Islam. June 1986.

44. Reported Masculinity Ratio in Pakistan: A Triumph of Anthropology and Economics over Biology.
Karol J. Krótki, June 1986.

45. Model Tables of Working Life.
R. Chow, P. Krishnan, and N.M. Lalu. July 1986.

46. A Macrosocial Analysis of Native Indian Fertility in Canada: 1961, 1971 and 1981.
Frank Trovato. September 1986.

47. The Family in Canada: A Support System for the Elderly.
Vijaya Krishnan. October 1987.

48. Young People's Explanations of Unemployment.
Graham S. Lowe, Harvey Krahn, and Julian Tanner. October 1987.

49. Gains in Life Expectancy in Canada on Elimination of Certain Causes of Death, 1978-1983.
Eddie Ng and P. Krishnan. November 1987.